Применение методов лингвистической семантики и машинного обучения для повышения точности и полноты поиска в поисковой машине «Exactus»

Applying linguistic semantics and machine learning methods to search precision improvement in search engine «Exactus»

Тихомиров И. А. (matandra@isa.ru), **Смирнов И. В.** (ivs@isa.ru)

Институт системного анализа РАН

Доклад посвящен проблемам применения методов лингвистической семантики и машинного обучения в задачах поиска и анализа текстов. Приведена экспериментальная оценка алгоритма поисковой машины Exactus в рамках семинара РОМИП. Сделаны выводы о перспективах использования логических методов машинного обучения в задачах анализа текстов.

1. Введение

В 2008 году поисковый алгоритм Exactus претерпел значительные изменения по сравнению с 2007 годом. Основной новинкой явилось включение в алгоритм анализа текстов контекстных правил установления значений минимальных синтактико-семантических единиц текста (синтаксем) [1]. Эти правила были получены с помощью методов машинного обучения на основе электронной версии синтаксического словаря Золотовой Г. А.[2]. Это позволило улучшить качество семантического анализа и снизить шум при поиске, что подтверждено результатами экспериментов в рамках российского семинара по оценке методов информационного поиска (РОМИП) [3]. О примененных методах машинного обучения и лингвистической семантики пойдет речь в данной работе.

2. Особенности поискового алгоритма Exactus

В различных публикациях неоднократно отмечалось, что поисковый алгоритм Exactus объединяет статистические и лингвистические методы поиска [4]. Лингвистическая составляющая поискового алгоритма заключается в учете семантики — смысловых значений слов, которые определяются на основании теории коммуникативной грамматики русского языка [5], [6].

Семантический анализ текста имеет своей целью извлечение смысла из текста и отображение его в формальную модель, которая позволяет находить смысловую близость двух текстов (применительно к задаче поиска — близость запроса и документа). При компьютерном семантическом анализе текста множество синтаксем каждого предложения отображается в неоднородную семантическую сеть, предложенную Г. С. Осиповым [7], с синтаксемами в вершинах и семантическими связями на множестве синтаксем в качестве ребер [8].

Семантический анализ текста оперирует в основном именными синтаксемами, которые выделяются в результате морфологического и синтаксического анализа. Именная синтаксема представляется в тексте именной или предложной группой — словосочетанием с существительным или предлогом в качестве управляющего слова. Именная синтаксема характеризуется морфологической формой — предлогом, падежом, и категориально-семантическим классом существительного, от которого она образована. Морфологическая форма синтаксемы и категориально-семантический класс определяются с помощью лингвистического анализатора текста. Синтаксема характеризуется также синтаксической функцией, которую она может выполнять в предложении, и синтаксическим значением. В ходе семантического анализа текста необ-

ходимо установить **значения** именных синтаксем, которые являются обозначениями смыслов, передаваемых текстом.

Морфологическая форма и категориально-семантический класс именной синтаксемы не однозначно задают её значение, а синтаксическую функцию, в которой выступает конкретная синтаксема, встречаемая в тексте в ходе анализа, автоматически определить невозможно. Таким образом существует проблема семантической многозначности синтаксем. Обычно для разрешения этой многозначности в анализ вовлекается контекст — глагол или отглагольное существительное, т.е. предикатное слово, при котором именная синтаксема встречается в предложении. Учет такого рода контекста требует создания специального словаря, описывающего наиболее частые сочетания определенного глагола с возможными синтаксемами при нем, и такой словарь был создан для глаголов и отглагольных существительных, наиболее часто встречаемых в текстах определенной тематики.

Словарь предикатных слов не может охватить все глаголы и отглагольные существительные, т.к. перечисление возможных синтаксем при глаголе вручную является весьма трудоёмкой задачей для лингвистов. Поэтому часто при семантическом анализе невозможно опираться на предикатное слово, так как его нет в словаре предикатных слов, следовательно, для точного установления значения синтаксемы в таких случаях необходимо учитывать другой контекст синтаксемы.

В безглагольных предложениях или предложениях, для которых предикатное слово не найдено в словаре, синтаксемы присутствуют рядом с другими элементами предложения, и несут своё значение только в данном контексте. Зависимость значения синтаксемы от собственных морфологических характеристик и характеристик соседних элементов предложения (не глаголов) является языковой закономерностью, которую необходимо обнаружить и зафиксировать для выполнения семантического анализа безглагольных предложений в дальнейшем. Такую закономерность для значения синтаксемы можно записать в виде **правила**, где в посылке правила находятся характеристики самой синтаксемы и контекста — окружающих её синтаксем и других элементов предложения, а заключение правила содержит значение, которое необходимо приписать целевой, рассматриваемой синтаксеме. Контекстные правила позволяют однозначно устанавливать значения синтаксем, т. е. снимать семантическую многозначность синтаксем.

Построение правил установления значений синтаксем экспертом-лингвистом требует больших трудозатрат на просмотр текстов, где встречаются анализируемые синтаксемы, анализ контекста синтаксем, обобщение признаков, влияющих на значение синтаксемы в разных текстах. Поэтому встала задача автоматического построения таких правил с помощью методов машинного обучения.

ДСМ-метод порождения гипотез, предложенный В. К. Финном [9], применяется для выявления скрытых причинно-следственных закономерностей в некоторой предметной области. Его задачей является обнаружение причин возникновения некоторого явления, или наличия свойств у объектов из некоторого множества. Решение этой задачи основывается на фактах или обучающем множестве объектов. Найденные причины используются для прогнозирования наблюдения явлений в дальнейшем.

Индуктивный вывод в ДСМ-методе основывается на принципе единственного сходства, сформулированном Д. С. Миллем: е*сли какое-то обстоятельство постоянно предшествует наступлению исследуемого явления, в то время как иные обстоятельства изменяются, то это обстоятельство есть, вероятно, причина данного явления.*

На практике построение гипотез о причинах того, что некоторые объекты обладают определенным свойством, заключается в нахождении характеристики сходства этих объектов — максимального множества признаков, которое принадлежит двум или более объектам с данным свойством. Эта характеристика сходства будет тем самым обстоятельством, которое не меняется от случая к случаю при наблюдении явления.

Для применения ДСМ-метода к решению задачи автоматического получения правил установления значений синтаксем было сделано следующее:

- введено отношение выводимости на морфологических признаках синтаксем (предлог, падеж, категориальный класс) и задана операция вычисления сходства морфологических признаков синтаксем, что позволяет вычислять характеристики сходства синтаксем;
- введено понятие составного морфологического признака — совокупности других морфологических признаков, рассматриваемых как один целый признак, задана операция вычисления сходства составных морфологических признаков;
- введено понятие синтаксем в контекстах, задана операция вычисления сходства синтаксем в контекстах.

Всё это позволило оперировать сложными лингвистическими объектами «синтаксема» или «синтаксема в контексте» без нарушения их внутренней логической структуры.

На описанных выше принципах была разработана программная реализация метода порождения правил установления значений синтаксем.

Материалом для построения обучающих примеров (синтакскем в контекстах) послужила электронная версия синтаксического словаря Г. А. Золотовой [2], предоставленная сотрудниками Машинного фонда русского языка Института русского языка РАН. В словаре приводятся синтаксемы с при-

мерами их вхождений в тексты литературы и периодики. В электронной версии словаря границы синтаксем выделены с помощью знаков подчеркивания «_». Это дает возможность автоматически выделить фрагменты текста, содержащие примеры синтаксем, и построить синтаксемы в контекстах. Следует заметить, что в синтаксическом словаре для каждой синтаксемы приводится очень мало примеров встречаемости в текстах, что делает неэффективным применение статистических методов машинного обучения. Преимущество логических методов машинного обучения состоит в том, что они срабатывают на обучающих выборках малого объёма (например, для индуктивного вывода в ДСМ-методе достаточно двух обучающих примеров).

В результате выполнения программной реализации было порождено более тысячи правил установления значений синтаксем. Для каждого правила сохранялись примеры, из которых оно было получено. Каждый пример содержит тексты, из которых были созданы целевая и соседняя синтаксемы, а также обрабатываемое предложение целиком. Таким образом, каждое правило хранит своё обоснование, которое может быть полезным как для оценивания адекватности реализации метода, так и при анализе лингвистом правильности установления значений синтаксем в дальнейшем.

Приведем пример правила установления значения «дестинатив» (назначение предмета или действия) для синтаксемы родительного падежа с предлогом «для»:

Правило: *Если встречается синтаксема в падеже <родительный> с предлогом <для>, имеющая категориальный класс <личное>, а до неё встречается синтаксема в падеже <именительный>, имеющая категориальный класс <предметное>, то полагается, что первая синтаксема имеет значение <дестинатив — назначение предмета или действия >*

Обоснование:

(40) ЗНАЧЕНИЕ = дестинатив
ЦЕЛЕВАЯ СИНТАКСЕМА = для тебя; КСК: личное
СОСЕДНЯЯ СИНТАКСЕМА = Все; ПРЕДЛОГ: ;ПАДЕЖ: им.вин.; КСК: предметное; ПОЗИЦИЯ: до
===КОНТЕКСТ: и песни, и силы — Все для тебя.

(41) ЗНАЧЕНИЕ = дестинатив
ЦЕЛЕВАЯ СИНТАКСЕМА = для различных рачков; КСК: личное
СОСЕДНЯЯ СИНТАКСЕМА = пища; ПРЕДЛОГ: ;ПАДЕЖ: им.; КСК: предметное; ПОЗИЦИЯ: до
===КОНТЕКСТ: Эти растения — пища для различных рачков

В примерах поле «КОНТЕКСТ» содержит предложение, из которого был построен пример.

Предложен алгоритм снятия смысловой многозначности синтаксем на основе полученных контекстных правил, который позволяет выбрать одно значение для синтаксемы из всех возможных, что уменьшает число ошибок семантического анализа текста в среднем в 4 раза (в случае срабатывания правила).

Реализованный алгоритм снятия смысловой многозначности синтаксем на основе полученных правил был внедрён в поисковый алгоритм Exactus, что позволило повысить точность семантического анализа и, соответственно, поиска документов.

3. Результаты экспериментов

Экспериментальная оценка поискового алгоритма Exactus, включающего методы лингвистической семантики и машинного обучения, проводилась в рамках российского семинара по оценке методов информационного поиска в 2008 году [10].

Общепринятым критерием оценки работы поисковых алгоритмов является 11-точечный график TREC, который отображает совмещенные показатели точности и полноты при разных показателях точности. На рисунках 1 и 2 ниже приведены 11-точечные графики TREC оценок AND и OR для системы Exactus и других участников семинара для коллекции «Белорусский WEB».

Заметим, что около 37% запросов, оцениваемых асессорами РОМИП, содержат синтаксемы с ролями, и только 16% из этих запросов содержат предикатные слова, т. е. в остальных запросах для синтаксем происходило снятие многозначности. Например, для двух оцениваемых запросов: «работа для студентов», «магазины для беременных» выполняется приведенное выше правило.

Разработчиками Exactus были сданы два прогона, которые отличались друг от друга настроечными параметрами поискового алгоритма. По графикам видно, что экспериментальный алгоритм Exactus получил ощутимо лучшие оценки по всем точкам TREC-графика для OR-оценки и большинству точек TREC-графика для AND-оценки.

4. Заключение

Применение методов интеллектуального анализа данных к обработке текстов на естественном языке может быть полезно для решения многих задач компьютерной лингвистики, но накладывает определённые требования на используемые методы

и получаемые результаты. Такие требования состоят, например, в способности оперировать сложными лингвистическими объектами и в интерпретируемости результатов. Систематическое применение в прикладной лингвистике методов машинного обучения, реализующих индукцию и аналогию, подтверждает возможность автоматического получения корректных результатов, которые наглядно объясняют некоторые языковые закономерности, а также помогают решать прикладные задачи, в частности снимать семантическую многозначность.

Массовость эксперимента, проведенного в рамках РОМИП, не позволяет оценить вклад отдельно каждого из используемых подходов к поиску. Анализ результатов участия в РОМИП показывает, что нельзя выделить какой-либо один фактор, существенно влияющий на показанное преимущество поискового алгоритма Exactus по сравнению с аналогами. Хороших результатов позволила достичь совокупность методов машинного обучения, лингвистической семантики, статистики, а также хороший уровень технического обеспечения и программирования.

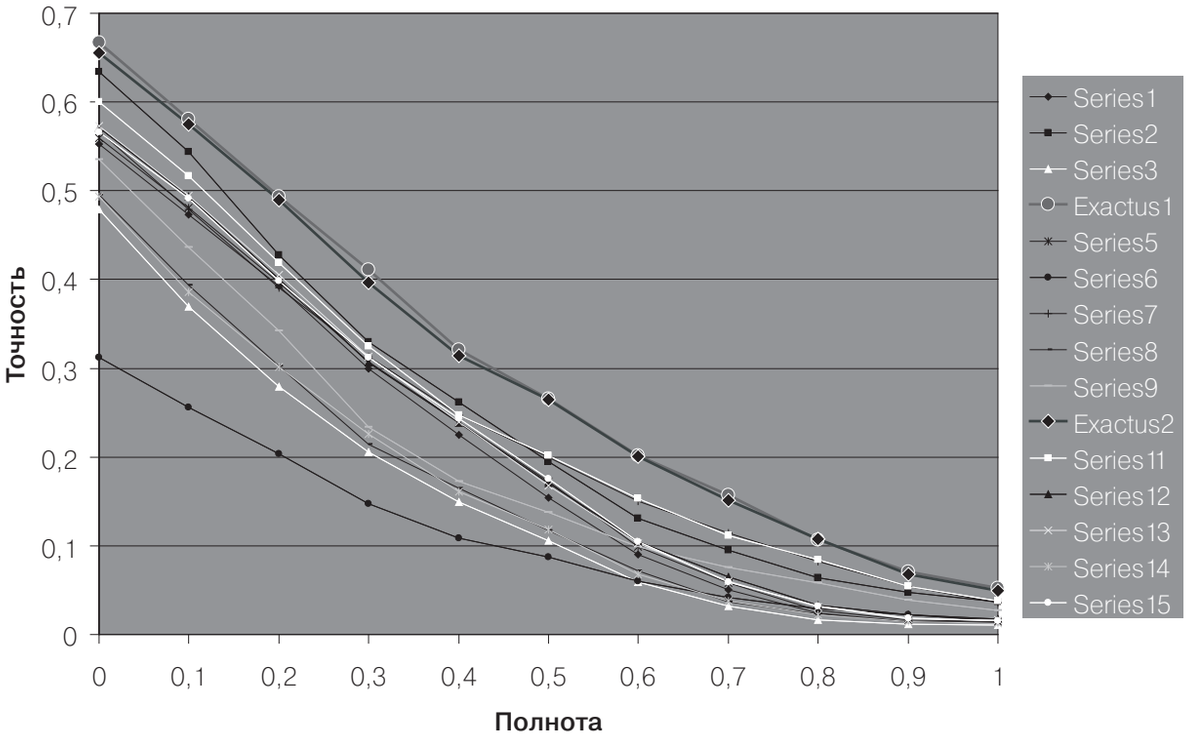

Рис. 1. Белорусский WEB, график TREC: OR-оценка.

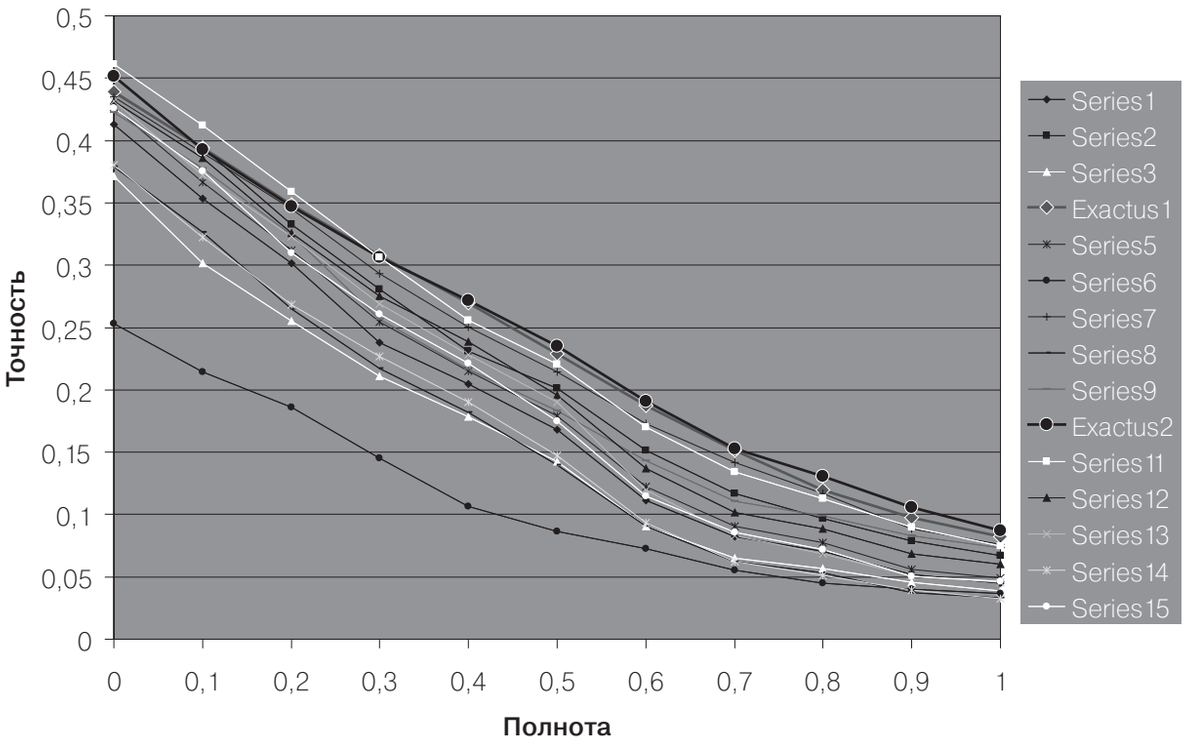

Рис. 2. Белорусский WEB, график TREC: AND-оценка.

Литература

1. *Смирнов И. В.* Метод автоматического установления значений минимальных синтаксических единиц текста // Информационные технологии и вычислительные системы. — 2008. — №3. — С. 30–45.
2. *Золотова Г. А.* Синтаксический словарь. Репертуар элементарных единиц русского синтаксиса // М.: Эдиториал УРСС, 2001.
3. *Смирнов И. В., Соченков И. В., Муравьев В. В., Тихомиров И. А.* Результаты и перспективы поискового алгоритма Exactus. // Труды российского семинара по оценке методов информационного поиска РОМИП'2007–2008. Санкт-Петербург: НУ ЦСИ. — 2008. — С. 66–76.
4. *Золотова Г. А., Онипенко Н. К., Сидорова М. Ю.* Коммуникативная грамматика русского языка // М.: Институт русского языка РАН им. В. В. Виноградова, 2004.
5. *Тихомиров И. А., Смирнов И. В.* Интеграция лингвистических и статистических методов поиска в поисковой машине Exactus. // Труды мезденародной конференции Диалог'2008. — С. 485–491.
6. *Gennady Osipov, Ivan Smirnov, Ilya Tikhomirov, Olga Zavjalova.* Application of Linguistic Knowledge to Search Precision Improvement.//Proceedings of 4th International IEEE conference on Intelligent Systems 2008. Volume 2. — P. 17-2–17-5.
7. *Осипов Г. С.* Приобретение знаний интеллектуальными системами // М.: Наука. Физматлит, 1997.
8. *Осипов Г. С., Смирнов И. В., Тихомиров И. А.* Реляционно-ситуационный метод поиска и анализа текстов и его приложения. // Искусственный интеллект и принятие решений. — 2008 — №2 — С. 3–10.
9. *Финн В. К.* ДСМ-метод как средство анализа каузальных зависимостей в интеллектуальных системах. // Научно-техническая информация, Сер. 2, Информ. процессы и системы. — №11. — 2000 — С. 1–5.
10. *Russian* Information Retrieval Evaluation Seminar // http://romip.ru